基于贝叶斯分区数据挖掘的光纤网络异常分析算法

刘云朋[1]，霍晓丽[1]，刘智超[2*]

(1. 焦作大学 信息工程学院，河南 焦作 454000；
2. 长春理工大学 光电工程学院，吉林 长春 130000)

摘　要：光纤网络通信中异常信息的快速、准确识别是保证通信稳定的关键，随着光纤网络通信数据的激增，也成为了近年来的一个研究热点。文中结合异常信息识别算法的精度与收敛速度之间的制约机理，提出了基于贝叶斯分区数据挖掘的异常信息识别算法。首先，采用贝叶斯定量完成数据样本的特征分类，通过极大化分析修正先验概率；然后，依据异常信息的不同类型设置挖掘特征参数及概率化系数；最后，依据贝叶斯分区分别对样本数据进行具有针对性的数据挖掘。实验以光纤局域网的通信状态数据为样本，将该算法与人工神经网络算法和遗传算法的识别结果进行对比，计算了三种算法的识别正确率、收敛速度以及算法稳定性。该算法的识别正确率均值为 93.83%，在数据量增大时未发生明显的降低。收敛速度与遗传算法相近，均值为 3.25 s。漏检率和误检率均值分别为 0.10% 和 0.54%。结果表明：该算法识别正确率与收敛速度均得到了提高，稳定性好，并能够在漏检率与误检率之间通过参数控制进行微调，具有较好的应用价值。

关键词：光纤网络；　异常信息识别；　数据挖掘；　贝叶斯分区

中图分类号：TP311　　**文献标志码：**A　　**DOI：**10.3788/IRLA20210121

Optical fiber network anomaly analysis algorithm based on Bayesian partition data mining

Liu Yunpeng[1], Huo Xiaoli[1], Liu Zhichao[2*]

(1. College of Information Engineering, Jiaozuo University, Jiaozuo 454000, China;
2. School of Optoelectronic Engineering, Changchun University of Science and Technology, Changchun 130000, China)

Abstract: The rapid and accurate identification of abnormal information in optical fiber network communication was the key to ensuring the stability of communication. The surge in conversion of optical fiber network communication data has also become the only research hotspot. Firstly, Bayesian partition data mining was used to quantify the feature classification of data samples, and the prior probability was corrected through maximization analysis; Secondly， the mining characteristic parameter and probability coefficient were set according to different types abnormal information; Finally, according to the Bayesian partition, the sample data was collected with specific data. The experiment takes the communication state data of the optical fiber interconnection as a sample, compared the recognition results of this algorithm with the artificial neural network algorithm and the genetic algorithm, and calculated the recognition accuracy, convergence speed and algorithm stability of the three algorithms. The average value of the recognition accuracy of this algorithm was converted to

收稿日期：2021−04−06；　**修订日期：**2021−05−12

基金项目：国家自然科学基金 (61703056)；吉林省优秀青年人才基金 (20190103154JH)

作者简介：刘云朋，男，副教授，硕士，主要从事计算机技术应用方面的研究。

通讯作者：刘智超，男，副教授，博士，主要从事光纤传感技术、光谱分析等方面的研究。

93.83%, and there was no significant decrease when the amount of data increased. The convergence speed was similar to that of genetic algorithm, with an average value of 3.25 s. The mean values of missed detection rate and false detection rate were 0.10% and 0.54%, respectively. The results show that the recognition accuracy and convergence speed of this algorithm are improved, the stability is good, and the parameter control can be fine-tuned between the missed detection rate and the false detection rate, which has better application value.

Key words: optical fiber network; abnormal information identification; data mining; Bayesian partition

0 引 言

光纤网络具有传输数据量大、交互节点多等特点，并广泛应用于通信领域，而随着客户端的不断增多以及原有设备的磨损老化，会出现断路、串联、跳线等错误，随之而来产生异常数据[1−2]。为了提高光纤网络通信的稳定性，对光纤网络中存在的异常信息进行快速识别具有重要意义。

光纤网络中异常信息的产生往往是由于设备故障或通信数据冲突造成的[3]，故其输出数据具有明显的特征，只要能够在海量的网络传输状态数据中进行快速分类，就能完成对异常位置、类型及其数据量的分析。对异常信息的识别，首先要从当前的信息中将错误信息的特征、类别进行先验分析，从而为异常信息的判别提供初始依据，再通过分析算法完成不同类型数据状态信息的判断。这个过程要将异常信息的特征与识别模型中的特征进行概率化匹配，从而完成对光纤网络中异常信息的精准识别。光纤网络状态信息监测算法有很多，诸如人工神经网络 (Artificial Neural Network, ANN)[4−5]、遗传算法 (Genetic Algorithm, GA)[6−7]、数据挖掘 (Data Mining, DM)[8−10] 等。人工神经网络的自主学习能力强，通过自适应分类可以对无序数据有效分类，具有很好的普适性，但对于数据量巨大的光纤网络数据，其容易产生局部最优的问题；遗传算法是通过模拟自然进化寻找最优解的，对于多元问题具有很好的适用性，对相似的光纤异常信息具有更好的区分性，但其遗传过程中必须携带一定量的上一代信息，这样会对异常信息的分类造成偏向性，影响新类型异常信息的识别。数据挖掘包括了多种数据分类方法，实际上是一种综合的数据分类手段，并且可以与不同的数据处理算法相结合，具有更高的兼容性。文中就是通过贝叶斯分区对数据进行预处理，再通过数据挖掘的手段进行分区识别，由此达到测试结果最优化的目的。

1 光纤网络的异常信息分类

1.1 贝叶斯分类原则

贝叶斯分类[11] 的核心思想是通过已知概率分布中存在的误判损失去完成数据分类的最优化。基于贝叶斯定理可知，在特征样本的条件下的类别概率 $P(K|X)$ 可以表示为：

$$P(K|X)=\frac{P(K)P(X|K)}{P(X)} \tag{1}$$

式中：K 表示类别；X 表示样本特征；$P(K)$ 表示可以预先获取的先验概率；$P(X|K)$ 表示关于样本特征的类别概率；$P(X)$ 表示算法设置系数。由上式可知将对类别概率的计算转化成了对先验概率与特征因子的计算。

1.2 异常信息分类算法设计

在光纤网络中，异常数据往往是具有一定特征的，并且产生的异常数据形式具有一定的相关性，从而采用先验概率去识别异常信息是有一定优势的。而在光纤网络中的异常信息往往是由于网络中错误代码、数据冲突等造成的，这些异常基本上是独立存在的，故在文中采用朴素贝叶斯策略[12] 进行分区，具有稳定性强、准确度高等特性。在这里想要计算类别概率时，所对应的类别就是算法的控制变量 k，由于控制变量并不唯一，故采用下标区分不同控制变量，m 个控制变量 k，即 k_1、k_2、k_3、$\cdots k_m$，其对应的特征矢量分别为 k_1、k_2、k_3、$\cdots k_n$。设样本为 $X=\{x_1, x_2, x_3, \cdots, x_n\}$，包括了则对于 K_i 满足在此条件下，其贝叶斯分类概率可表示为：

$$P(K_i|X)=\frac{P(K_i)P(X|K_i)}{P(X)} \tag{2}$$

式中：$P(X)$ 为设置系数。故当公式 (2) 中分子满足极大化时，则该式也能够满足。在光纤网络传输的通常情况下控制变量的概率是由于硬件设备决定的，换言之从概率分布的角度而言，这个也往往被看作是常量。故最终实际上是在计算 $P(X|K_i)$ 的，则其可表达为：

$$P(X|K_i)=\prod_{k-1}^{n}P(X_k|K_i) \tag{3}$$

在计算样本数据时可以获取公式 (3) 中不同 X 赋值时的 $P(X|K_i)$, 故当其符合公式 (3) 时样本数据被分类到 K_i 中, 从而样本符合极大化要求。

针对样本的先验概率[13], 如果样本集合中所有的样本或训练集都没有出现某个分量值, 则检测结果为 0。并且采用拉氏平滑[14] 修正先验概率, 从而防止非特征数据占据特征数据类别的问题。如果训练样本 D 中类别量为 N, 则对应的第 i 个特征值对应数值为 N_i, 由此获得修正结果:

$$P(X|K_i)=(|D_c|+1)(|D_c|+N_i)^{-1} \tag{4}$$

由上式可知, 当样本总数增大时, 修正过程中的先验效应造成的影响会越来小, 其估计值与真实概率会无限逼近。

2 基于贝叶斯分区的异常信息数据挖掘算法

2.1 数据挖掘算法

在通过贝叶斯定理完成异常信息分区后, 对已完成分区的样本数据进行数据挖掘, 挖掘过程主要分为: 特征数据提取、数据预处理、分区分类、模型构建。首先, 对已完成的分区进行信息类型趋势分析, 从而对不同的异常信息的数据格式与类型进行分类; 然后, 对异常信息进行概率化处理, 将异常信息的概率属性叠加概率化系数上; 最后, 利用贝叶斯拓扑结构[15], 将概率化[16] 的数据分布转化为数据特征向量, 形成数据挖掘的边界条件。

设数据集合为 A, 挖掘特征参数为 B, 异常信息的分类系数为 n, 概率化系数为 l, 则数据挖掘的计算规律满足:

$$\int\bar{P}=\left\{P(a_n)\sum_{i\in n}^{B_i}l(A_i,B^n)\forall\right\} \tag{5}$$

为了提高数据挖掘的精度与挖掘速度之间的制约关系, 采用贝叶斯分区将初始海量光纤网络数据进行分区, 这样在数据挖掘过程中不同分区的侧重是不同的, 针对不同异常信息类型其概率化值不同 (该概率化系数可以理解为每个数据点的权值), 从而挖掘深度和速度可以达到最优化配置, 避免无效挖掘, 从而保证挖掘速度。

设任意贝叶斯分区中数据集合为 X, 而对应的 X 中可以展开成 $n\times n$ 的矩阵形式, 与第 1 节中的样本数据集对应, 则满足其分区数据挖掘的概率关系有:

$$X=\sum_{n\in N}\begin{pmatrix}x_{11} & x_{12} & \cdots & x_{1n}\\ x_{21} & x_{22} & \cdots & x_{2n}\\ \cdots & \cdots & \cdots & \cdots\\ x_{n1} & x_{n2} & \cdots & x_{nn}\end{pmatrix} \tag{6}$$

根据以上步骤完成迭代每一个贝叶斯分区中的数据集合, 就能快速地获得全部的异常数据集合。

2.2 算法实现

为了提高光纤网络中异常信息识别精度与收敛速度, 将贝叶斯分区应用于数据挖掘前的数据分区, 从而使不同分区中异常信息类型的识别概率可以根据分区属性进行调节, 这样就能提高异常信息的识别精度与收敛速度。挖掘算法的流程如图 1 所示, 实现步骤如下:

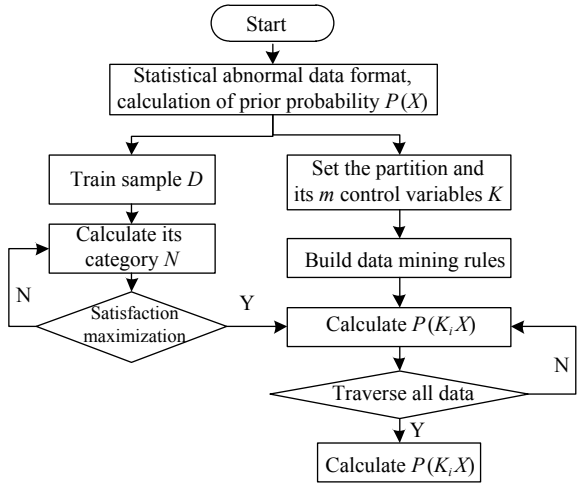


图 1 基于贝叶斯分区数据挖掘算法流程图

Fig.1 Flow chart of data mining algorithm based on Bayesian partition

(1) 对光纤网络中异常信息的种类与数据格式进行分类, 并根据以往异常信息出现频次的差异设定不同的先验概率 $P(X)$;

(2) 设置分区内样本数据集 $X=\{x_1, x_2, x_3, \cdots, x_n\}$, 依据异常信息特征设置 m 个控制变量 k, 即 k_1、k_2、k_3、$\cdots$、k_m;

(3) 循环判断符合控制变量条件下数据集的概率, 当其满足极大化条件时, 输出贝叶斯分类概率值 $P(K_i|X)$;

(4) 训练样本数据 D, 设置其需要处理的数据的类别量 N 和其对应数值 N_i, 从而对原有的贝叶斯分类

概率值进行修正，随着数据量不断增大，修正效果将无限逼近真实概率，从而提高系统分区精度，最终确定所以数据的区域划分；

(5) 在具有明确分区的基础上，将数据挖掘的计算规律给出，并将贝叶斯分区作为其边界条件，对不同区域的异常信息进行概率化分类，分类依据为公式(5)，对数据集合 A 中的 n 个类别进行挖掘；

(6) 通过分区数据挖掘的概率关系作为收敛条件对所有分区进行分段迭代，将光纤网络中数据遍历后输出异常信息结果。

3 对比实验

采用实验室内光纤局域网模拟光纤通信网络，计算机采用 32 位 Windows 10 系统，主频 3.0 GHz 双核处理器，内存 2.0 GB 为硬件基础。以网络延迟、光开关断路、数据信道占用率为主要标志参数，本算法数据挖掘语言采用 VS 平台 C++实现，数据服务器的处理器选用至强 E5 型。为了对比异常信息识别效果，针对相同的光纤通信数据，分别采用人工神经网络(Artificial Neural Network, ANN) 和遗传算法 (Genetic Algorithm, GA) 进行异常信息提取与识别。

3.1 识别正确率对比

首先对算法的识别正确率进行比较，比较的指标采用识别正确率 P 表示，识别正确率定义为判定为异常信息的数据样本为真的个数与判定数据样本的总数的比值，每个样本为光纤网络传输数据包，包含一个通信时刻所有的状态参数信息，则分析 1 000 个数据包样本的测试结果如图 2 所示。

由图 2 可知，在数据样本总量大幅增加的情况下，该算法的识别正确率基本保持不变，平均值为 93.83%，而 ANN 算法和 GA 算法的平均值分别为 83.34% 和 82.92%，可以看出该算法的识别正确率明显优于两种传统识别算法。

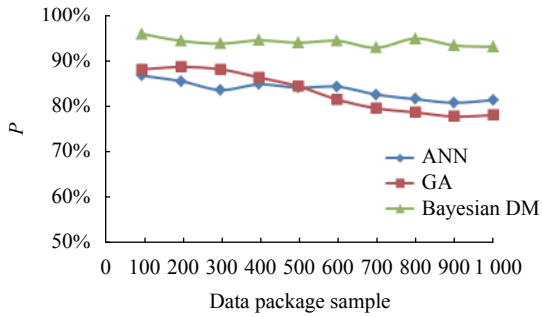


图 2 识别正确率对比

Fig.2 Comparison of recognition accuracy rate

3.2 算法处理速度分析

在保证识别精度的基础上，算法收敛速度就成为了判断算法优劣的第二个重要指标，同样将三种算法对同一组数据的处理时间进行比较，结果如图 3 所示。

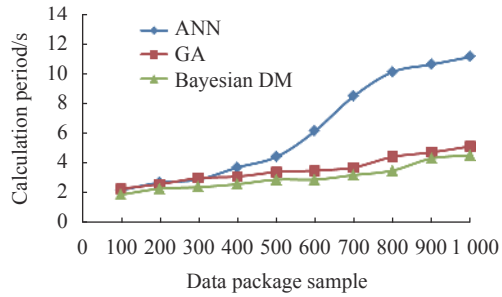


图 3 不同算法处理速度对比

Fig.3 Comparison of processing speed of different algorithms

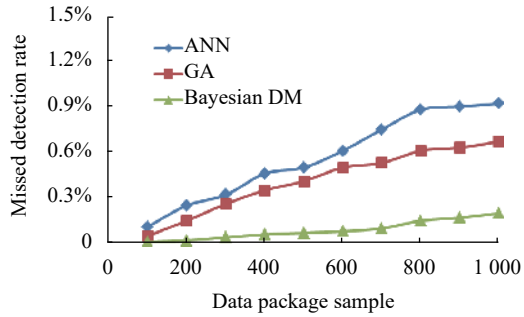
由测试结果可知，当数据样本小于 400 时，三种算法的收敛时间基本一致，当数据样本大于 400 后，ANN 算法的收敛时间明显增大，而 GA 算法与该算法的收敛时间相近。分析认为，由于这个值并不是单纯的 400 个点，而是 400 个数据单元，每个单元中处理数据信息还有状态参数的，当超过 400 时，其数据运算量就会显著增大，故计算周期由此发生较大差异。表明该算法测试速度方面符合设计要求。

3.3 算法可靠性分析

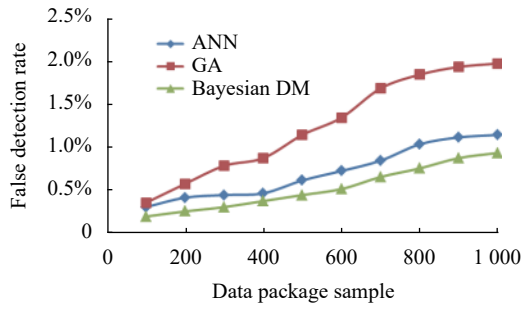
对算法可靠性分析主要从漏检率与误检率两个方面进行评价，漏检是将异常信息判定为正常信息，其比率为漏检信息数量与总检测信息数量的比值，误检是将正确信息错误判断成异常信息，其比率为误检信息数量与总检测信息数量的比值，对比三种算法的可靠性如图 4 所示。

由图 4(a) 可知，本算法的漏检率平均值为 0.10%，ANN 和 GA 算法的平均值为 0.58% 和 0.43%，在数据量增大时，算法的漏检率没有明显增大，具有较好的可靠性。由图 4(b) 可知，该算法的误检率平均值为 0.54%，ANN 和 GA 算法的平均值为 1.26% 和 0.72%，总体变化趋势平稳。分析认为误检率高于漏检率的原因是数据挖掘的控制因子设置较大，侧重全部检

出。若当实际情况要求尽量避免误检时，可以通过调小控制因子，使算法结果侧重避免错检。



(a) 漏检率
(a) Missed detection rate



(b) 误检率
(b) False detection rate

图 4 算法可靠性对比

Fig.4 Comparison of algorithms reliability

4 结 论

文中针对在光纤网络通信中异常信息的识别正确率与收敛速度之间的制约问题，提出了基于贝叶斯分区数据挖掘的异常信息识别算法。该算法将贝叶斯分区应用于数据样本分类，再通过分区数据挖掘实现异常信息的快速识别。实验将该算法与两种常用的识别分类算法进行比较，结果显示：该算法的识别正确率、收敛速度以及稳定性均具有一定优势，在光纤网络通信异常分析中具有一定的实用价值。

参考文献：

[1] Ramezani M, Yaghmaee F. A review on human action analysis in videos for retrieval applications [J]. ***Artificial Intelligence Review***, 2016, 46(4): 485-514.

[2] Wang Hui, Zhang Cuiyu. Differences between network data mining algorithm based on improved genetic algorithm [J]. ***Computer Simulation***, 2015, 32(5): 311-314. (in Chinese)

[3] Kuang Y, Guo Y, Xiong L, et al. Packaging and temperature compensation of fiber Bragg grating for strain sensing: A survey [J]. ***Photonic Sensors***, 2018, 8(4): 320-331.

[4] Jia Q. Location and monitoring of fiber optic line faults [J]. ***China New Telecommunications***, 2017, 19(1): 74-74.

[5] Yeung S, Russakovsky O, Jin N, et al. Every moment counts: dense detailed labeling of actions in complex videos [J]. ***International Journal of Computer Vision***, 2018, 126(24): 375-389.

[6] Chen Yang, Zhao Shanghong, Wang Xiang, et al. BER analysis of high-altitude OFDM-FSO modulation system under exponentiated weibull atmospheric turbulence model [J]. ***Laser & Infrared***, 2018, 48(7): 832-837.

[7] Chen Y, Li L J. Very fast decision tree classification algorithm based on Red-Black tree for data stream with continuous attributes [J]. ***Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition)***, 2017, 37(2): 86- 90.

[8] Liu Y, Wang C R. An improved big data clustering method based on sampling fusion [J]. ***Microelectronics & Computer***, 2017, 34(4): 17- 21.

[9] Gu X Q, Jiang Y Z, Wang S T. Zero-order TSK-type fuzzy system for imbalanced data classification [J]. ***Acta Automatica Sinica***, 2017, 43(10): 1773-1788.

[10] Lee J, Lee S, Hwang I. Hybrid system modeling and estimation for arrival time prediction in terminal airspace [J]. ***Journal of Guidance Control & Dynamics***, 2016, 39(4): 903-910.

[11] Sun B C, Li J Z, Zhang W T. Fiber Bragg grating sensor [J]. ***Optical Fiber Sensing and Structural Health Monitoring Technology***, 2019, 26(4): 77-148.

[12] Huang X, Wang Z, Li Y, et al. Design of fuzzy state feedback controller for robust stabilization of uncertain fractional-order chaotic systems [J]. ***Journal of the Franklin Institute***, 2015, 351(12): 5480-5493.

[13] Shang F, Yi J, Xiong A, et al. A node localization algorithm based on multi-granularity regional division and the lagrange multiplier method in wireless sensor networks [J]. ***Sensors***, 2016, 16(11): 1934.

[14] Pan Q K, Sang H Y, Duan J H, et al. An improved fruit fly optimization algorithm for continuous function optimization problems [J]. ***Knowledge-Based Systems***, 2014, 62(1): 69- 83.

[15] Guan L, Hu G J, Wang Zh. Research on network security situational awareness technology based on big data [J]. ***Netinfo Security***, 2016, 1(9): 45-50.

[16] Guo H, Liu H, Wu C, et al. Logistic discrimination based on G-mean and F-measure for imbalanced problem [J]. ***Journal of Intelligent and Fuzzy Systems***, 2016, 31 (3): 1155-1166.